Probabilistic Data Generation for Deduplication and Data Linkage

Peter Christen

Data Mining Group, Australian National University
Contact: peter.christen@anu.edu.au

Project web page: http://datamining.anu.edu.au/linkage.html

Funded by the ANU, the NSW Department of Health,
and the Australian Research Council (ARC) (LP #0453463)

- Data linkage and deduplication
- Data linkage techniques
- Test data for data linkage
- Artificial data
- Probabilistic data set generator
- Example data generated
- Experimental study
- Conclusions and outlook

Data linkage and deduplication

- The task of linking together records representing the same entity from one or more data sources (patient, customer, business, etc.)
- Real world data is *dirty*, so cleaning and standardisation is important
- Applications of data linkage
  - Remove duplicates in a data set (internal linkage)
  - Merge new records into a larger master data set
  - Create customer or patient oriented statistics
  - Compile data for longitudinal studies
  - Geocode data (match addresses with geographic reference data)

Data linkage techniques

- Computer assisted linkage goes back to 1950s
- Deterministic linkage
  - Exact linkage (if a *unique identifier* of high quality – precise, robust, stable over time – is available)
  - Rules based linkage (complex to build and maintain)
- Probabilistic linkage (*Fellegi & Sunter*, 1969)
  Apply linkage using available (personal) information (which can be missing, wrong, coded differently, or out-of-date)
- Modern approaches
  Based on machine learning, data mining, or information retrieval techniques (clustering, decision trees, active learning, learnable string metrics, graphical models, etc.)

Test data for data linkage

- Various data sets are used in recent publications (*restaurant*, *cora*, *citeseer*, *census*, etc.)
  - Usually very small (less than 2,000 records)
  - Proprietary and even confidential data has been used
- There is a lack of standard test data
- Hard to compare new algorithms and to learn how to use and customise data linkage systems
- Recent small repository: *RIDDLE*
  `http://www.cs.utexas.edu/users/ml/riddle/`
  (Repository of Information on Duplicate Detection, Record Linkage, and Identity Uncertainty)

Artificial data

- Privacy issues prohibit publication of real data (for example of names, addresses, dates of birth, etc.)
- De-identified or encrypted data cannot be used (as linkage algorithms work on name and address strings)
- Artificial data as alternative to real data
  - Based on real data (frequency and misspellings tables)
  - Must model content and statistical properties of real data
- Advantages
  - Content and error modifications can be controlled
  - Data can be published
  - Easy to repeat and verify experiments

A probabilistic data set generator

- First data generator by *Hernandez & Stolfo* (1996)
- Improved by *Bertolazzi et.al.* (2003) (no details given, not publicly available)
- Our generator
  - Open source (Python)
  - Part of the *Febrl* data linkage system (Freely extensible biomedical record linkage)
  - Easy to modify and improve by a user
  - Based on real world frequency look-up tables for names, addresses, date of birth, etc.
  - Includes look-up tables with real typographical errors

Data generation

- Step 1: Create original records
  Randomly select values from various frequency look-up tables, or from a user specified range (e.g. for *date of birth*)
- Step 2: Create duplicates based on original records by introducing modifications
  - Single errors (insert, delete, substitute a character; transpose two characters)
  - Insert or delete a whitespace (split or merge a word)
  - Set to missing (empty string), or insert new value
  - Swap with another value from a look-up table
  - Swap two attribute values (e.g. *given name* ↔ *surname*)

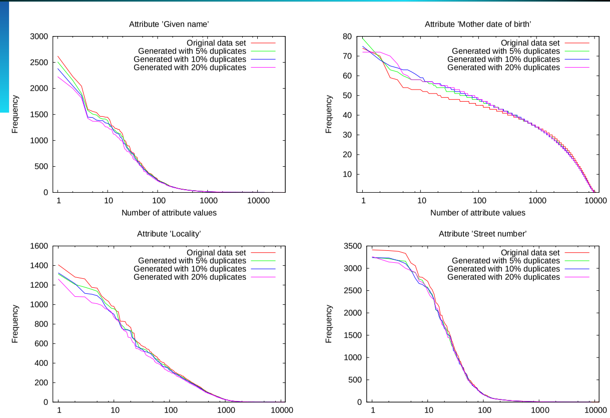
- Data set with 4 original and 6 duplicate records

```
      REC_ID,           ADDRESS1,          ADDRESS2,  SUBURB
   rec-0-org,        wylly place,     pine ret vill,   taree
 rec-0-dup-0,         wyllyplace,     pine ret vill,   taree
 rec-0-dup-1,     pine ret vill,        wylly place,   taree
 rec-0-dup-2,        wylly place,     pine ret vill,   tared
 rec-0-dup-3,       wylly parade,     pine ret vill,   taree
   rec-1-org,      stuart street,          hartford,  menton
   rec-2-org,   griffiths street,            myross,   kilda
 rec-2-dup-0,   griffith sstreet,            myross,   kilda
 rec-2-dup-1,    griffith street,           mycross,   kilda
   rec-3-org, ellenborough place, kalkite homestead,  sydney
```

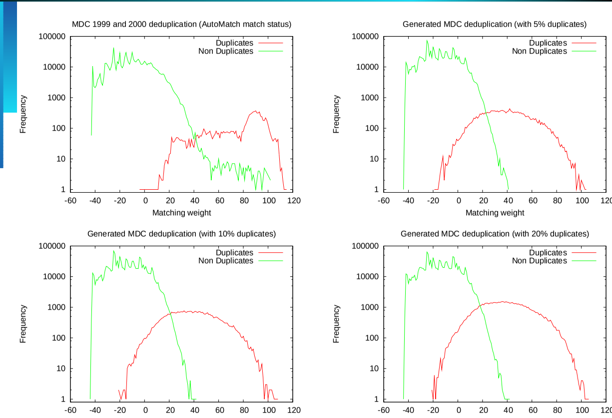
- Each record is given a unique identifier, which allows the evaluation of accuracy and error rates

- NSW Midwifes Data Collection (MDC)
  - Extracted years 1999 and 2000 (175,211 records)
  - Contained 5,331 twin and 177 triplet births
- Linkage done by *AutoMatch* resulted in 8,442 duplicate record pairs
- Extracted frequency tables for mother's name, address and date of birth attributes
- Created 3 data sets with 175,211 records each, containing 5%, 10%, and 20% duplicates
- Then performed deduplication using *Febrl* data linkage system

Sorted attribute frequencies



Deduplication matching weights



Conclusions and outlook

- Several possible improvements
  - Relax independence assumption (based on real world frequency tables), for example a change of address results in new street name, number and type, as well as postcode and locality
  - Allow generation of groups of records, for example for households (census)
  - Fine tune error modifications (scanning, typing, etc.)
- Do further comparison studies with real data sets
- See project web page for more information

  http://datamining.anu.edu.au/linkage.html